# TV program classification based on face and text processing

Wei, G. Agnihotri, L. Dimitrova, N.
Dept. of Comput. Sci., Wayne State Univ., Detroit, MI, USA;
*This paper appears in:* **Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on**

Meeting Date: 07/30/2000 - 08/02/2000
Publication Date: 30 July-2 Aug. 2000
Location: New York, NY USA
On page(s): 1345 - 1348 vol.3
Volume: 3
Reference Cited: 11
Number of Pages: 3 vol. xxxv+17778
Inspec Accession Number: 6704325

**Abstract:**
In this paper we describe a system to classify TV programs into predefined categories based on the analysis of their video contents. This is very useful in intelligent display and storage systems that can select channels and record or skip contents according to the consumer's preference. Distinguishable patterns exist in different categories of TV programs in terms of human faces and superimposed text. By applying face and text tracking to a number of training video segments, including commercials, news, sitcoms, and soaps, we have identified patterns within each category of TV programs in a predefined feature space that reflects the face and text characteristics of the video. A given video segment is projected to the feature space and compared against the distribution of known categories of TV programs. Domain-knowledge is used to help the classification. Encouraging results have been achieved so far in our initial experiments

**Index Terms:**
face recognition feature extraction image classification text analysis tracking video signal processing TV program classification channel selection content recording content skipping domain knowledge face processing face tracking human faces intelligent display systems intelligent storage systems superimposed text text processing text tracking training video segments video content analysis

**Documents that cite this document**
There are no citing documents available in IEEE Xplore at this time.

# TV PROGRAM CLASSIFICATION BASED ON FACE AND TEXT PROCESSING

Gang Wei[1], Lalitha Agnihotri[2], and Nevenka Dimitrova[2]
[1]Computer Science Department
Wayne State University, Detroit, MI 48202
[2]Philips Research
345 Scarborough Rd, Briarcliff Manor NY, 10510

## ABSTRACT

In this paper we describe a system to classify TV programs into predefined categories based on the analysis of their video contents. This is very useful in intelligent display and storage systems that can select channels and record or skip contents according to the consumer's preference. Distinguishable patterns exist in different categories of TV programs in terms of human faces and super-imposed text. By applying face and text tracking to a number of training video segments, including commercials, news, sitcoms, and soaps, we have identified patterns within each category of TV programs in a predefined feature space that reflects the face and text characteristics of the video. A given video segment is projected to the feature space and compared against the distribution of known categories of TV programs. Domain-knowledge is used to help the classification. Encouraging results have been achieved so far in our initial experiments.

## 1. INTRODUCTION

Consumers today are receiving increased number of channels while their disposable time for TV viewing is declining as revealed by a study in [3]. Viewers normally switch among dozens of channels to search for programs that fit their taste. Consumers often time-shift programs of interest and then selectively watch segments of interest while fast forwarding through commercial breaks. Content selection is becoming increasingly important in this context and new methods are required to offer features on the future display and storage devices.

Despite various proposals on the features of the next generation TV/VCR systems, we believe that the most essential one is to provide intelligent services to the consumers, including the creation of a personalized TV profile, channel selection support and content searching/filtering based on consumer's preference as mentioned in [5]. This requires the ability to classify the TV programs into predefined categories based on the analysis of the video and/or audio content. In addition, a system for the segmentation of commercials from news programs is described in [6].

In this paper we propose a scheme to classify given video segments into one of four categories of TV programs, namely **news, commercial, sitcom,** and **soap** by tracking face and super-imposed text. This stems from the fact that most TV programs focus on human activities and face detection is important in identifying program category. In most occasions video content features can be sufficiently extracted by capturing face movements. On the other hand, text is a helpful cue in recognizing certain types of TV programs. We observed that there exist distinguishable pattern for each category of TV programs in terms of the occurrence of faces and text. By integrating information about faces and text, the system classifies a given video segment into one of the above categories assuming that it belongs to one of the four categories. It can be extended to recognize more categories by adding new classification rules.

The program categorization consists of two phases, namely training and classification. In the training phase, we use a number of video segments to construct a feature space, where the dimensions correspond to the face and text information in the video segments. The distribution patterns of different categories of TV programs are modeled in this space. In the classification phase, a given video segment is projected on to the feature space. The features used for classification are derived from tracking of faces and the super-imposed text in the video stream. The probability that a segment belongs to one of the four categories is evaluated by a weighted distance metric combined with domain-knowledge. The segment is classified to belong to the category with the highest probability.

The remaining part of this paper is organized as follows. The description of the feature extraction phase is given in Section 2. This includes a brief overview of our face and text tracking algorithms. Section 3 presents the training and classification of patterns within each TV program category and the use of domain-knowledge to enhance the accuracy of classification. In Section 4 we present the experimental results, and Section 5 concludes the paper by discussing the possible future extensions of the system.

## 2. FEATURE EXTRACTION

The features used for classification are derived from the tracking of faces and of super-imposed text in the video stream. The results of face and text tracking are the basis for later classification. Therefore the tracking capability has the primary responsibility for the

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

performance of the whole system. Object tracking involves two issues, namely the detection of the targets in each frame and the extraction of object trajectories over frame sequences to capture their movements.

### 2.1 Face tracking

Various methods have been proposed for the detection of human faces [1, 9]. A comprehensive survey can be found in [4]. In our system we employed the scheme described in [11] with some slight modification in choosing the YUV color coordinate instead of YIQ for skin-tone region segmentation to adapt to the MPEG-1 and MPEG-2 framework. This system is accurate owing to the utilization of different features and the novel iterative partition process. However, due to the complexity of computation, applying face detection to each frame is inefficient. To improve the speed of face tracking, we took advantage of the content continuity between consecutive frames by considering the joint detection of faces and trajectory extraction. The variation of faces within a continuous shot is usually small. Compressed-domain cut detection is performed to segment the video segment into shots [8]. Face detection is then applied to the first few frames of each shot. For each detected face the mean and standard deviation in color, height, width and center position are computed. All these features constitute a face model. The face model is used for tracking faces in the future frames till the next cut is detected.

In the following frames the tracking system just searches in a reduced area in the neighborhood of the face found according to the model created in the previous frame for the corresponding face instead of searching the whole frame. The models are updated on a frame by frame basis to reflect the latest changes until the end of the shot. Our experiments showed that by reducing the search space we gain over ten times speed-up against applying face detection to each isolated frame.

### 2.2 Text tracking

Tracking of visual text is treated differently in that its detection and trajectory extraction are two independent processes. This is because of two major factors. First, the text detection algorithms usually work very fast and there is no great speed-up if we consider inter-frame continuity within shots. Second, unlike faces, super-imposed text may survive scene changes (cuts) and therefore tracking can not be performed for separate shots. The text detection as proposed in [2] is applied to each frame of the video. Other techniques used for detecting text can be found in [7,10]. Detected text boxes in consecutive frames are compared with each other and merged as belonging to the same trajectory if they are similar enough. The tracking of text and face are integrated. Fig. 1 shows the results of this on two frames, where faces and texts are enclosed with white rectangles.

## 3. TRAINING AND CLASSIFICATION OF VIDEO

The program categorization consists of two phases, namely training and classification. Section 3.1 discusses the training of the models and Section 3.2 discusses the classification based on this training information.

### 3.1 Training

The presence, size, and continuity patterns of faces and text are different for each category of TV programs. This fact is also reflected in the face/text distribution graphs in Fig. 2, each of which is the 3-D plot of a typical video segment belonging into news, commercial, sitcom, and soap. The x, y, z coordinates in the 3-D plot correspond to row, column, and frame number respectively. A filled (empty) circle at position (*row, column, frame_number*) indicates the presence of a face (text) box centered at (*row, column*) in frame *frame_number*. We can see that commercials often contain a lot of text and few faces. In contrast, text and faces often occur at the same time in news. In soaps and sitcoms text regions are rare, and usually there are many close-up face shots lasting for a long time in soaps while in sitcoms faces have smaller size and shorter duration

To identify distinguishable characteristics of different categories of TV programs, we have constructed a feature space based on face and text tracking results. Number (per unit time) and average duration of face and text trajectories are dimensions in this feature space. Further, faces and texts with long duration or close shot size are more important in recognizing TV programs. The face and text trajectories are filtered by duration and shot size threshold. The number and average duration of the "survived" trajectories constitute additional dimensions in the feature space. In addition, the count and duration of face trajectories with faces larger than shoulder shots are also dimensions in the feature space.

A number of video segments containing the four categories of TV programs are used as the training set. Face and text tracking is applied, converting them into vectors in the feature space. The center of each category is computed by averaging the vectors.

### 3.2 Classification

To classify a given video segment, we map it into the same feature space and evaluate its probability of being each category by the weighted distances to the centers of the news, commercial, sitcom and soap clusters. The weights of the dimensions in distance computation are set empirically to maximize the separability between different categories of TV programs.

Besides face and text trajectories, the use of domain knowledge can help the classification. For example, in news and commercials, optical (gradual) cuts are frequently used by the editor to ensure a more smooth and pleasant visual effect while in sitcoms and soap most cuts between shots are abrupt. Thus, the percentage of optical cuts among all cuts can be used as an extra feature to update the probability value. When a video segment contains a high portion of optical cuts, a choice of news or commercial is favored over sitcom or soap. The experiment shows that the utilization of domain knowledge improves the accuracy by 20%. The video segment is classified as the category with the highest probability.

## 4. EXPERIMENTAL RESULTS

We used 26 video segments exclusively for the training set, which includes five segments for news, six for commercials, eleven sitcoms and four soaps. Face and text tracking is performed using I-frames in the MPEG segments. The testing set contains 35 segments different from the training set. Out of these, 27 segments are correctly labeled while eight are misclassified. Therefore, an accuracy of 77.1% was achieved. Most errors originate from news segments labeled as commercials and sitcoms labeled as soaps. New programs like "Datelines" in NBC contain fewer faces than regular news programs, which makes them seem like commercial as far as face and text trajectories are concerned.

## 5. CONCLUSION AND FUTURE WORK

A TV program classification system based on face and text tracking is described in this paper. The main contribution of this research work is the exploration of the face and text patterns in different categories of TV programs. Domain knowledge is applied to aid the classification process, which improves the accuracy. The system is flexible and extendable in that once we obtain new observations and more complete domain knowledge, new rules can be added to enable the system to recognized more categories of TV programs. So far we have achieved encouraging results and better performance is expected by further optimization.

However the combination of feature space and domain knowledge has its drawbacks. The effort of watching many video segments to find out new domain knowledge is tedious. Moreover, adding new classification rules will eventually result in a very complicated system. We are developing a learning system based on Hidden Markov Models, which will elegantly solve this problem by automatically learning the new patterns (rules). Another direction of future extension is the utilization of audio content, which is especially powerful in discriminating sitcoms and soaps by detecting laughter.

## 6. ACKNOWLEDGEMENTS

Thanks are due to Dr. Ishwar K. Sethi for the contribution of face detection algorithm and Dongge Li for MDC, the MPEG decoder.

**Reference:**

1. M. Abdel-Mottaleb and Ahmed Elgammal, "Face Detection in Complex Environments from Color Images," *Proc. ICIP,* Kobe Japan 1999.
2. L. Agnihotri and N. Dimitrova. Text Detection in Video Segments. *Proc. of Workshop on Content Based Access to Image and Video Libraries*, pp 109-113, June 1999.
3. J. Consoli, "The 11.8-Hour Daily Diet", *Mediaweek*, Vol.8, No. 16, April, 98, pp9-12
4. R. Chellappa, C.L. Wilson and S. Sirohey, "Human and machine recognition of faces: A Survey", *Proceedings of the IEEE*, Vol. 83, No. 5, pp. 705-740, 1995.
5. N. Dimitrova, "Digital Video for the Time Impaired", *IEEE Multimedia*, April-June, 1999, pp14-17
6. A.G. Hauptmann and M.J. Witbrock. Story Segmentation and Detection of Commercials in Broadcast News Video. Proceedings of *Advances in Digital Libraries Conference*, Santa Barbara, CA., April 22-24, 1998
7. R. Lienhart and F. Stuber, "Automatic Text Recognition for Video Indexing," SPIE conference on image and video processing, Jan 1996.
8. N.V. Patel and I.K. Sethi. Compressed Video Processing for Cut Detection. *IEE Proceedings: Vision, Image and Signal Processing*, Vol.143, pp.315-323, Oct. 1996.
9. H. Rowley, S. Baluja, T. Kanade, "Neural Network-based face detection," *Computer Vision and Pattern Recognition*, San Francisco, 1996.
10. J.-C. Shim, C. Dorai, and R. Bolle, "Automatic Text Extraction from Video for Content-Based Annotation and Retrieval," In Proc. of the *International Conference on Pattern Recognition*, pp. 618-620, 1998.
11. G. Wei and I.K. Sethi. "Face Detection for Image Annotation", *Pattern Recognition Letters*, Vol. 20, 1999, pp. 1313-1321.

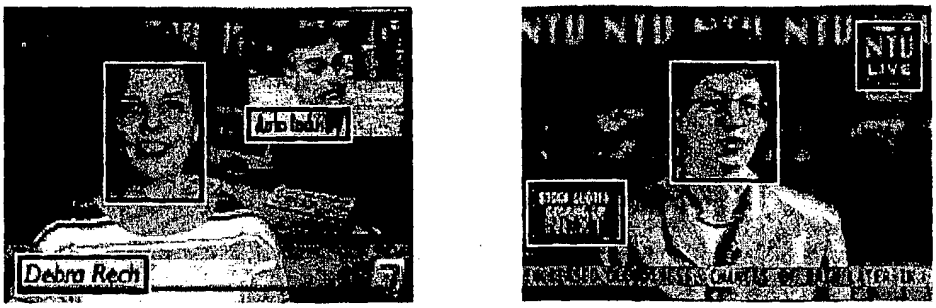
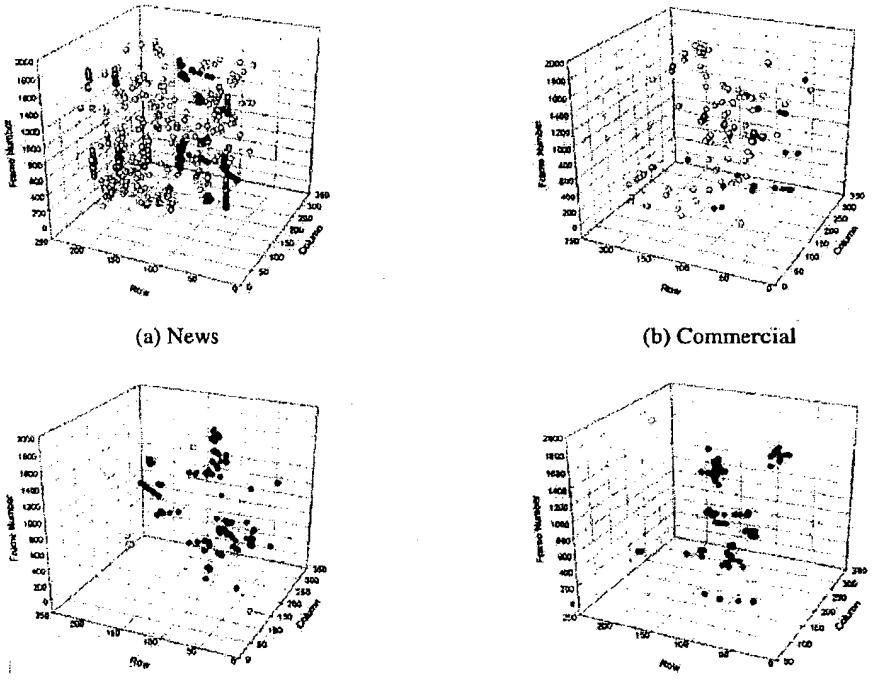



Fig. 1 Examples of integrated face and text detection



(a) News

(b) Commercial

(c) Sitcom

(d) Soap Opera

Fig. 2 3D Plot of face and text trajectories in news, commercial break, situational comedy and soap opera segments. Filled circles stand for faces and blank circles stand for text. Row and column of the center of the text/face bounding box are along x and y axes respectively while frame number is along the z-axis.